Perturbative methods for ESS census tables

Keywords: SDC, disclosure, perturbative, table, census

1. INTRODUCTION

The population census has been an important output of official statistics for a long time. In the current global situation, researchers are increasingly interested in combining census information from different countries. In Europe an attempt to facilitate this more easily was taken by Eurostat in developing the Census Hub: a software system where census tables from the participating EU member states can be found. These member states each filled the Census Hub with their own census tables. The census tables are about personal information and are giving information at quite detailed level: crossings of many descriptive variables, often with detailed categories. Hence, even though tabulated data, some statistical disclosure control methods are needed to protect the individual privacy of the people in the census tables of those member states. Up until the 2011 census, all member states defined and used their own 'proper' SDC method(s). This led to the undesired situation that although the tables from the member states were available through a single portal, combining the information over different countries in a useful way was in some cases difficult if not impossible.

Consequently, a project[1] in which the European Centre of Excellence on SDC was asked to harmonise the SDC approaches of the different countries was launched. As a result two SDC methods were proposed: targeted record swapping (TRS) and adding random noise based on the cell-key method (CKM). These methods were tested by the partners of the Centre of Excellence on SDC that took part in the project, using SAS software obtained from ONS.

By suggesting to use one or both of these methods as harmonised approach by the EU member states, it became apparent that a more general implementation in (existing) open source SDC software would be needed. Hence, Eurostat asked the Centre of Excellence on SDC to provide these implementations.

In the current extended abstract, we will quickly describe the suggested methods and introduce the implementations that have been developed. At the time of writing this abstract the tests of the new implementations were ongoing, hence results of those were not yet available. However, at the NTTS meeting in 2019 we will be able to show test results of all project partners and hopefully also of some voluntary testers from outside the project.

2. METHODS

For the protection of the hypercubes of the European 2021 census, two methods are proposed by the Centre of Excellence on SDC: targeted record swapping (TRS) and random noise based on the cell-key method (CKM). In this section we will briefly describe the

[1] The project is partly funded by the EU project 'Open Source tools for perturbative confidentiality methods', Specific Grant Agreement N° 2018.0108 under the Framework Partnership Agreement N° 11112.2014.005-2014.533.

two methods. See the deliverables of that project for more detailed information: https://ec.europa.eu/eurostat/cros/content/harmonised-protection-census-data_en.

2.1. Targeted record swapping

Targeted record swapping (TRS) is a pre-tabular method, applied to the microdata that underlie the census frequency count tables. Swapping means that certain scores on certain variables will be interchanged between certain records. For example, assume two records i and j are selected to have their scores x_i and x_j respectively on variable X swapped. Then record i will get score x_j and record j will get score x_i assigned, The main characteristics of the TRS approach for the census are:

- TRS swaps households, i.e., individuals are not swapped separately.
- For each level of geographic hierarchy, households of high disclosure risk are determined.
- Only the geographical variables will be swapped between pairs of 'similar' households.
- Swaps are made in every geographic hierarchy level.
- Households which do not fulfil *k*-anonymity are swapped.
- At the lowest hierarchical level an additional number of households is swapped, such that in total a minimum number of households defined by a pre-set percentage will be swapped.

2.2. Cell-key method

Random noise based on the cell-key method (CKM) is a post-tabular method: random noise will be added to each table cell, according to some noise probability distribution and a mechanism to draw from that noise distribution. Essentially, the method considered to be used for the census hypercubes is a slightly modified version of the method proposed in [1]. CKM assigns random *record* keys to the records in the microdata underlying the census frequency count tables. Those record keys are drawn from a uniform U(0,1) distribution. When constructing a cell in a census table, not only the number of records in that cell is counted, but their record keys are added as well and the fractional part of the resulting number is set to be the *cell* key. The cell key can thus be regarded as a uniform U(0,1) variable as well. Given pre-set transition probabilities (the so-called *p*-table), the noise is then determined as function of the *cell* key and the cell *value*. This way consistency is guaranteed: whenever the same cell appears in some table, the cell value as well as the cell key will be the same and thus the added noise will be the same.

To construct a *p*-table, an R-package was developed. This package is made available via https://github.com/sdcTools/ptable.

Additivity is not guaranteed when applying CKM based random noise. This is due to the fact that the noise will be added to each cell individually, i.e., additivity constraints between inner cells and marginal cells are not taken into account when adding the noise. When proposing this method, the Centre of Excellence got the impression from the census working group that consistency would be valued higher as opposed to the additivity.

3. SOFTWARE

For the above two mentioned methods, Open Source software implementations are developed. The software is available at the http://github.com/sdcTools website. We will briefly describe the available software.

3.1. Targeted Record Swapping

For TRS a library was programmed in C/C++ and the sources are available at the repository recordSwapping on github. The general call to perform the record swapping is given by

```
RecordSwap(std::vector<std::vector<int>> data,
           std::vector<int> similar, std::vector<int> hierarchy
           std::vector<int> risk, int hid, int th,
           double swaprate, int seed)
```

where `data` is the rectangular table containing the microdata with only integers as scores, `similar` is a vector of indices of the variables in `data` that define similarity of households, `hierarchy` is a vector of indices of the variables in `data` that define the geographic hierarchy, `risk` is a vector of indices of the variables in `data` that are used to determine the *k*-anonymity risk, `hid` is the index of the household identifier in `data`, `th` is an integer defining the *k* of the *k*-anonymity, `swaprate` is a number between 0 and 1 defining the minimum proportion of households to be swapped and seed is an integer defining the `seed` for the random number generator.

The library can be used from different user interfaces, using appropriate wrappers. E.g., the Rcpp package can be used to import the functionality into R. We have also made TRS available through the µ-Argus program.

3.2. Cell Key Method

CKM is essentially a method that comprises of data wrangling and table building. Therefore, two implementations are made available: an implementation in R as the cellKey package and an implementation in τ-Argus.

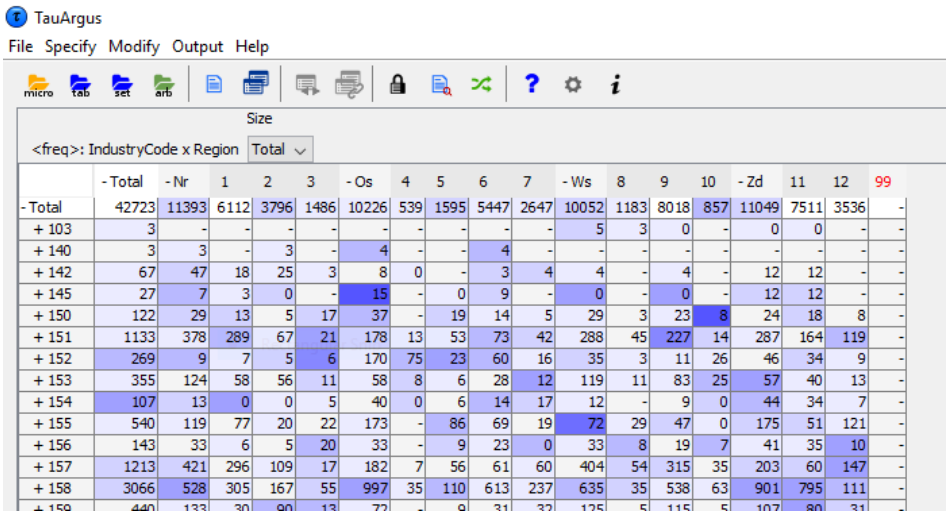
3.2.1. cellKey package

The cellKey package is already slightly more general than actually needed for the census table protection. It can deal with frequency count tables using the original method of the ABS as well as the slightly adjusted CKM. Moreover, it can deal with (weighted) magnitude tables and weighted frequency tables using the method as described in [1]. Moreover, it can generate record-keys in several ways. In this abstract we will focus on the use for frequency count tables of the 2021 census, using the slightly adjusted CKM and assume that the record-keys are added to the microdata as an additional variable, generated from a uniform U(0,1) distribution. With the cellKey package you can specify the table and get the perturbed table using CKM. For the perturbation a *p*-table has to be specified. Potential *p*-tables can be generated with the ptable R-package which is a dependency of the cellKey package.

You can get information on the applied perturbations by using the `print` and `summary` commands.

3.2.2. T-Argus implementation

With the τ-Argus implementation, we again assume that the record-keys are added to the microdata as a separate variable. The *p*-table to be used to determine the perturbations can be generated by the ptable R-package. The file containing the *p*-table can be specified in the metadata section of τ-Argus. The implementation can also deal with weighted frequency count tables using the method as described in [1]. Once you have specified the frequency count table, you can choose the Cell Key Method as one of the 'Suppress' options. After application of the method, the perturbations can be made visible using the 'Colored view' (see Figure 1 for an example): the darker the color, the larger the applied noise.



	- Total	- Nr	1	2	3	- Os	4	5	6	7	- Ws	8	9	10	- Zd	11	12	99
- Total	42723	11393	6112	3796	1486	10226	539	1595	5447	2647	10052	1183	8018	857	11049	7511	3536	-
+ 103	3	-	-	-	-	-	-	-	-	-	5	3	0	-	0	0	-	-
+ 140	3	3	-	3	-	4	-	-	4	-	-	-	-	-	-	-	-	-
+ 142	67	47	18	25	3	8	0	-	3	4	4	-	4	-	12	12	-	-
+ 145	27	7	3	0	-	15	-	0	9	-	0	-	0	-	12	12	-	-
+ 150	122	29	13	5	17	37	-	19	14	5	29	3	23	8	24	18	8	-
+ 151	1133	378	289	67	21	178	13	53	73	42	288	45	227	14	287	164	119	-
+ 152	269	9	7	5	6	170	75	23	60	16	35	3	11	26	46	34	9	-
+ 153	355	124	58	56	11	58	8	6	28	12	119	11	83	25	57	40	13	-
+ 154	107	13	0	0	5	40	0	6	14	17	12	-	9	0	44	34	7	-
+ 155	540	119	77	20	22	173	-	86	69	19	72	29	47	0	175	51	121	-
+ 156	143	33	6	5	20	33	-	9	23	0	33	8	19	7	41	35	10	-
+ 157	1213	421	296	109	17	182	7	56	61	60	404	54	315	35	203	60	147	-
+ 158	3066	528	305	167	55	997	35	110	613	237	635	35	538	63	901	795	111	-

Figure 1. Colored view of a table in τ-Argus protected by CKM based random noise

The table can be saved in the standard ways of τ-Argus, but also in a new ckm-format. That format optionally contains the original cell value, the perturbed cell value, the difference as well as the cell-keys. Additional information on the perturbations is saved in the report file produced by τ-Argus whenever a table is saved.

4. CONCLUSIONS

The Centre of Excellence on SDC has produced implementations in Open Source SDC software to apply the TRS and the CKM based random noise to the 2021 census data. This way it has become easier for member states to apply and test the methods to their own data. Moreover, in case they decide to use one or both methods, they can now start to think about incorporating any one of the implementations into their production process.

At the time of writing this abstract the tests of the new implementations were ongoing, hence results of those were not yet available. However, at the time of the NTTS 2019 we will be able to show the first results from applications of the implementations to real census data.

REFERENCES

[1] G. Thompson, S. Broadfoot and D. Elazar, Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics, paper at the 2013 Work Session on Statistical Data Confidentiality www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf